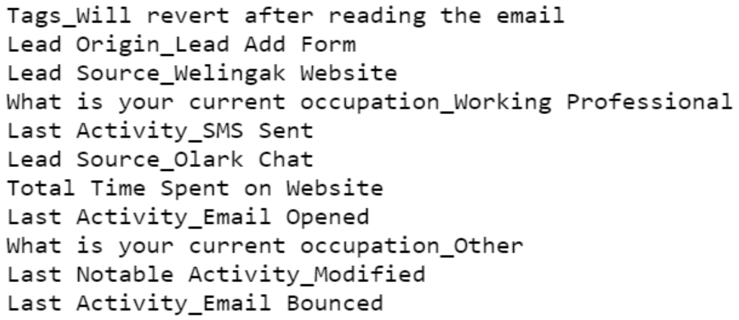
Summary Report

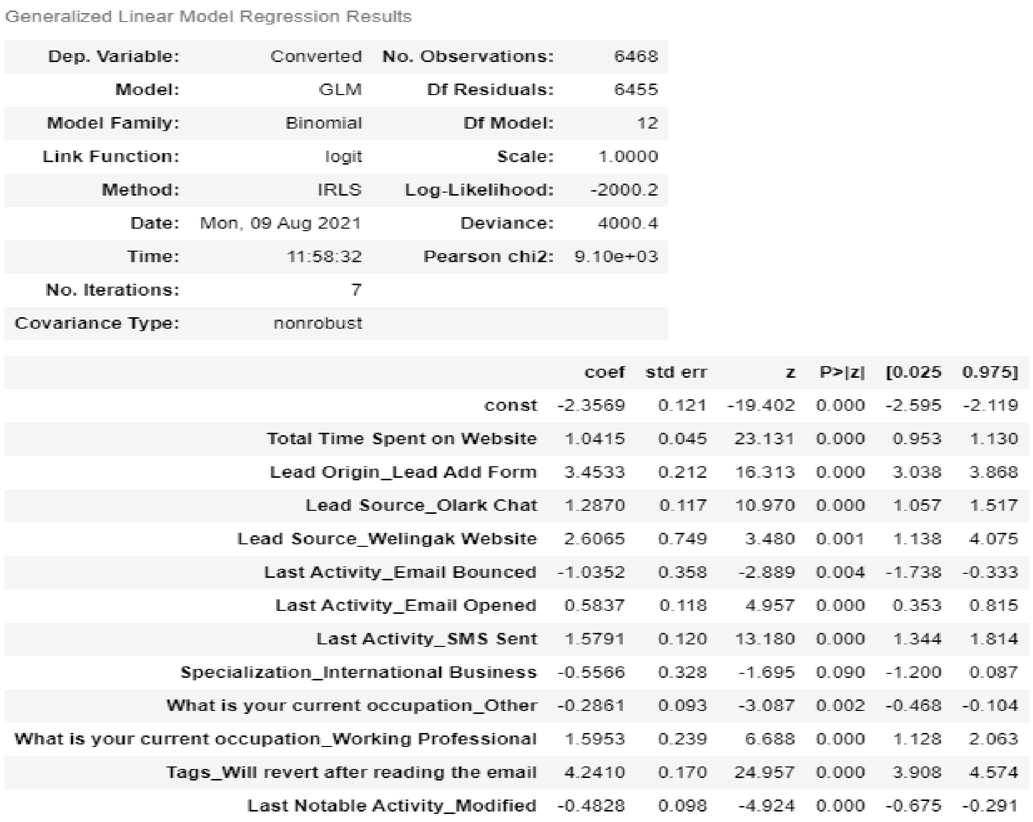
Assignment Approach:

1. Basic data check and data cleaning steps and ensuring to retain 70% of data
    a. Checking the shape, data types, duplicates, stats of numerical features
    b. Converting all ‘Select’ values to null
    c. Null handling
    d. Outlier treatment
    e. Handling features with skewed data or with too many categories
2. EDA to get a first view of the data trend, correlation with the target variable and within predictors
3. Model Building Steps:
    a. Conversion from ‘Yes/No’ values to 0/1
    b. Dummy creation for categorical features
    c. Train-test split (70-30 ratio)
    d. Scaling the numerical features
    e. Selecting top 20 features by using RFE process
    f. Building 1st model using statsmodel’s GLM method, verified the summary report to check P-value and then validated the VIF
    g. Dropped the features having P-value more than .05 and/or VIF more than 5 one by one and rebuilding the model. This process went in iteration and 10 iterations was done to arrive at final model where all the features had P-value and VIF within acceptance. In this process below features were retained:



```
Tags_Will revert after reading the email
Lead Origin_Lead Add Form
Lead Source_Welingak Website
What is your current occupation_Working Professional
Last Activity_SMS Sent
Lead Source_Olark Chat
Total Time Spent on Website
Last Activity_Email Opened
What is your current occupation_Other
Last Notable Activity_Modified
Last Activity_Email Bounced
```

    h. Below is the model summary report

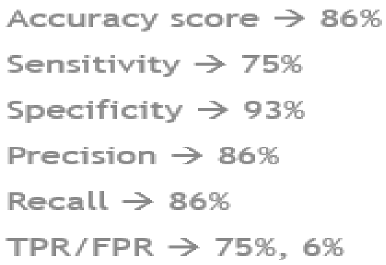


Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6455
Model Family:	Binomial	Df Model:	12
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2000.2
Date:	Mon, 09 Aug 2021	Deviance:	4000.4
Time:	11:58:32	Pearson chi2:	9.10e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.3569	0.121	-19.402	0.000	-2.595	-2.119
Total Time Spent on Website	1.0415	0.045	23.131	0.000	0.953	1.130
Lead Origin_Lead Add Form	3.4533	0.212	16.313	0.000	3.038	3.868
Lead Source_Olark Chat	1.2870	0.117	10.970	0.000	1.057	1.517
Lead Source_Welingak Website	2.6065	0.749	3.480	0.001	1.138	4.075
Last Activity_Email Bounced	-1.0352	0.358	-2.889	0.004	-1.738	-0.333
Last Activity_Email Opened	0.5837	0.118	4.957	0.000	0.353	0.815
Last Activity_SMS Sent	1.5791	0.120	13.180	0.000	1.344	1.814
Specialization_International Business	-0.5566	0.328	-1.695	0.090	-1.200	0.087
What is your current occupation_Other	-0.2861	0.093	-3.087	0.002	-0.468	-0.104
What is your current occupation_Working Professional	1.5953	0.239	6.688	0.000	1.128	2.063
Tags_Will revert after reading the email	4.2410	0.170	24.957	0.000	3.908	4.574
Last Notable Activity_Modified	-0.4828	0.098	-4.924	0.000	-0.675	-0.291

4. Prediction step → conversion probability was predicted on the final model using the X_train data set on final list of features. Initial conversion indicator was predicted based on default threshold of 0.5 i.e. if the probability is more then 0.5 then 1 else 0. 1 means converted and 0 means not converted
5. Model was evaluated with the help of below shown Metrics:



Accuracy score → 86%

Sensitivity → 75%

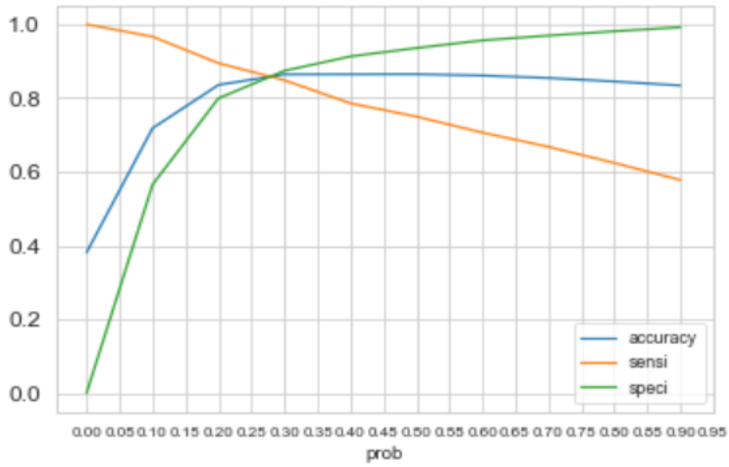
Specificity → 93%

Precision → 86%

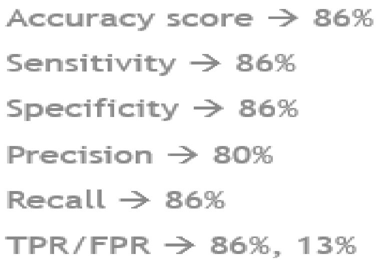
Recall → 86%

TPR/FPR → 75%, 6%

6. Optimal cut-off (0.28) was found by plotting the accuracy, sensitivity and specificity as shown below:



7. Steps 6, 7 repeated with new cut-off to get below metrics:



Accuracy score → 86%

Sensitivity → 86%

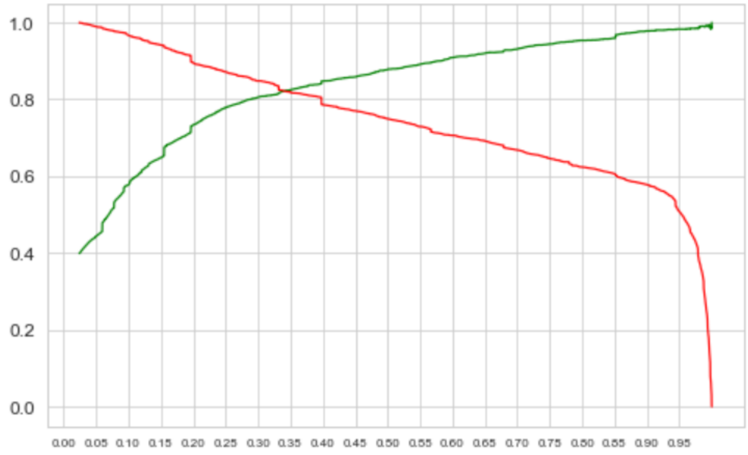
Specificity → 86%

Precision → 80%

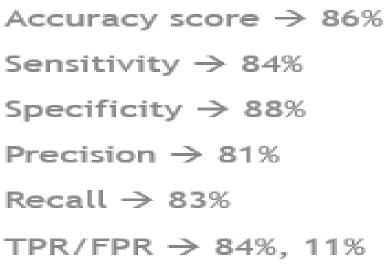
Recall → 86%

TPR/FPR → 86%, 13%

8. Precision and Recall trade-off was plotted as shown below to get the new optimal cut-off (0.33) which is to be used on test data set:



9. Steps 6, 7 repeated to get below metrics:



Accuracy score → 86%

Sensitivity → 84%

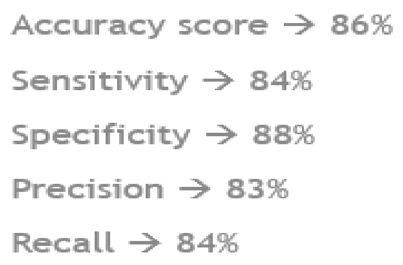
Specificity → 88%

Precision → 81%

Recall → 83%

TPR/FPR → 84%, 11%

10. Prediction was made on test data set using the final model and new cut-off (0.33), the metrics on test data as follows:



Accuracy score → 86%

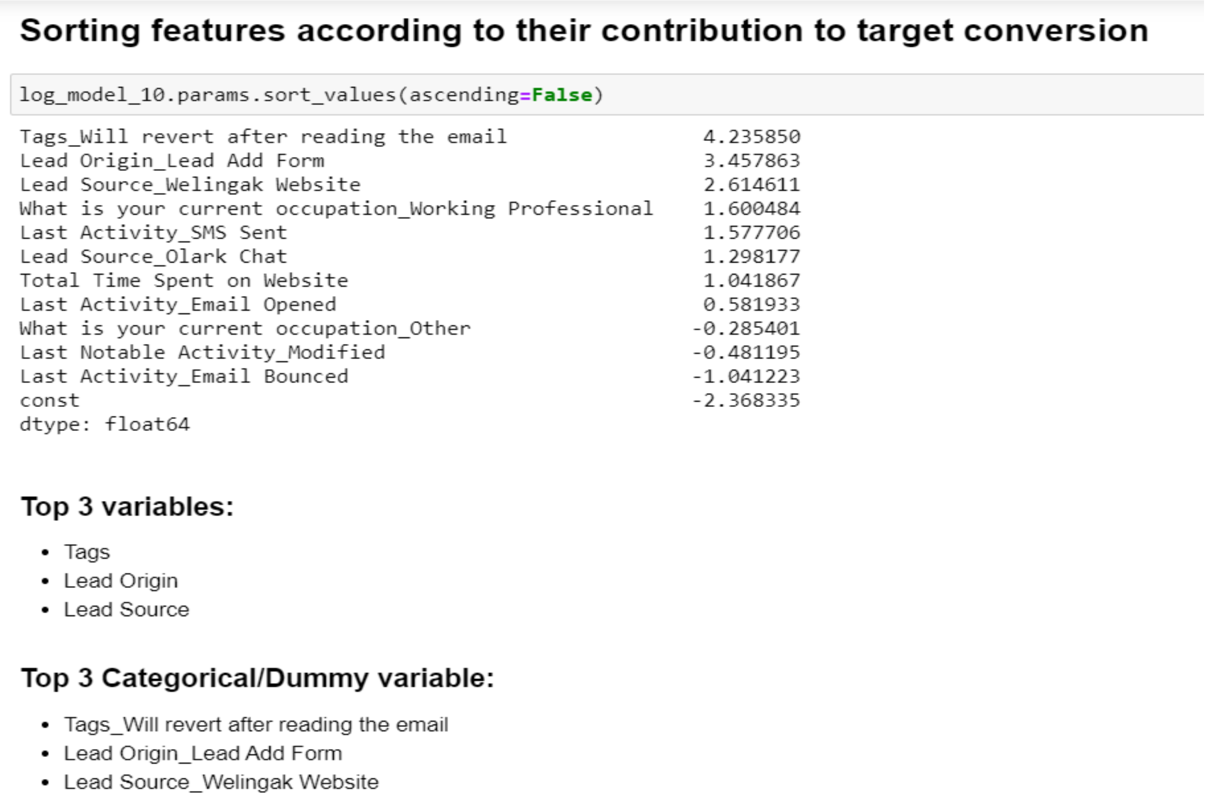
Sensitivity → 84%

Specificity → 88%

Precision → 83%

Recall → 84%

11. Lead scoring was done by multiplying 100 to the conversion probability.
12. The final set of categories impacting the conversion rate are as shown below:



Sorting features according to their contribution to target conversion

```
log_model_10.params.sort_values(ascending=False)
```

```
Tags_Will revert after reading the email                 4.235850
Lead Origin_Lead Add Form                                3.457863
Lead Source_Welingak Website                             2.614611
What is your current occupation_Working Professional     1.600484
Last Activity_SMS Sent                                   1.577706
Lead Source_Olark Chat                                   1.298177
Total Time Spent on Website                              1.041867
Last Activity_Email Opened                               0.581933
What is your current occupation_Other                   -0.285401
Last Notable Activity_Modified                          -0.481195
Last Activity_Email Bounced                             -1.041223
const                                                   -2.368335
dtype: float64
```

Top 3 variables:

- Tags
- Lead Origin
- Lead Source

Top 3 Categorical/Dummy variable:

- Tags_Will revert after reading the email
- Lead Origin_Lead Add Form
- Lead Source_Welingak Website

Learning from the process:

1. Problem faced and learning
    a. Some columns had values as ‘Select’. This value is same as ‘Null’ hence needed handling
    b. There were many features having so many less frequent categories to them making the data very skewed. Hence to have better model, we had to group less-frequent categories to a new category.
    c. To overcome data imbalance issue, we had to drop some features having very skewed data or having only 1 category.
    d. Finally, we got 55 features which made it impossible to manually select the initial set of features. Hence, we used RFE to get the top 20 features and then used manual model building steps for further feature elimination

2. Recommendation - Sales team should target below customer to maximize the conversion rate:

a. Customers tagged as ‘will revert after reading the email’
b. Someone who is filling the form
c. Working professionals
d. If someone sends SMS
e. Someone spending more time on the website